DOCUMENT RESUME

ED 477 929 TM 035 025

AUTHOR	Shermis, Mark D.; Raymat, Marylou Vallina; Barrera, Felicia
TITLE	Assessing Writing through the Curriculum with Automated Essay Scoring.
SPONS AGENCY	Fund for the Improvement of Postsecondary Education (ED), Washington, DC.
PUB DATE	2003-04-00
NOTE	30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
CONTRACT	PR116B000387A
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*College Students; *Essays; Higher Education; *Portfolio Assessment; Portfolios (Background Materials); Scoring; *Test Scoring Machines; *Writing Evaluation; Writing Improvement

ABSTRACT

This paper provides an overview of some recent work in automated essay scoring that focuses on writing improvement at the postsecondary level. The paper illustrates the Vantage Intellimetric (tm) automated essay scorer that is being used as part of a Fund for the Improvement of Postsecondary Education (FIPSE) project that uses technology to grade electronic portfolios. The purpose of the electronic portfolio is to demonstrate a mechanism for translating the general learning goal on writing in an operational way that permits the developmental tracking of students throughout their undergraduate curriculum. Moreover, the technology can be readily incorporated into any course in which writing is a significant component. (Contains 22 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Assessing Writing through the Curriculum with Automated Essay Scoring

Mark D. Shermis

Marylou Vallina Raymat

Felicia Barrera

Educational and Psychological Studies

Florida International University

ED 477 929

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. Shermis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM035025

Paper presented at the Annual Meetings of the American Educational Research Association, Chicago, IL. Support for this work was supplied, in part, by the Fund for the Improvement of Post-Secondary Education (PR116B000387A). The views or opinions expressed in this paper are the authors and do not necessarily reflect those of FIPSE or the U.S. Department of Education.

BEST COPY AVAILABLE

Abstract

This paper is designed to provide an overview of some recent work in automated essay scoring that focuses on writing improvement at the post-secondary level. We intend to illustrate the Vantage Intellimetric™ automated essay scorer that is being used as part of a FIPSE project which employs the technology to grade electronic portfolios. The purpose of the electronic portfolio is to demonstrate a mechanism for translating the general learning goal on writing in an operational way that permits the developmental tracking of students throughout their undergraduate curriculum. Moreover, the technology can be readily incorporated into any course where writing is a significant component.

Introduction

This paper is designed as a second update on our progress with automated essay scoring in evaluating writing at the post-secondary level (cf. Shermis & Daniels, 2001, Shermis & Barrera, 2002; Shermis, in press). This effort has been funded through a FIPSE grant and is focused on providing feedback for evaluating papers that might be included in an electronic portfolio. The research and development for this grant is designed to address one small aspect of the larger problem: How do we assess undergraduate general education, or as they are sometimes called, "principles of undergraduate learning"?

Most institutions will typically identify between six and nine dimensions of general education or a similar number of undergraduate learning principles. For example, almost every institution has something regarding students' ability to "reason quantitatively" or to "respect diversity". Invariably one of these principles will be: "the ability to communicate effectively". The good news is that there will generally be a base of agreement among these principles—we are impressed with how readily they replicate from one institution to the next.

The bad news is that there are significant disagreements as how to operationalize what is meant by the various constructs, and typically there are competing definitions as to how one, for

example, "communicates effectively". One need only travel to an assessment conference and stroll the exhibit areas to see the vast array of options being marketed by competing vendors.

The measurement of "communicating effectively" can take a number of forms ranging from administering an objective test to evaluating student documents written in a capstone seminar. Criticisms of assessment techniques currently used are typically aimed at characteristics such as the incorporation of restricted (departmental or unit-wide norms), insufficient or lack of information about validity and reliability, reliance on idiosyncratic rubrics, and failure to identify factors contributing to student growth in progressing throughout the program (Shermis & Barrera, 2002). While there has been no bona fide sentiment to standardize on one approach, techniques that would permit cross-institutional comparisons have been in demand over the past twenty years.

A measurement procedure that holds some promise in overcoming these difficulties is the electronic portfolio. Similar to typical portfolios, it is a purposeful organization of learner-selected evidence of school and non-school accomplishments, but stored on electronic media including floppy disks, CD-ROMs, or the World Wide Web (Stemmer, 1993). The definition has several important components. First, the phrase "purposeful organization" suggests that the "evidence" contained

in the portfolio constitutes something more than a "grab bag" of materials. Usually the work represents the best example of what the learner is capable of doing for a particular class of products. For example, a psychology major might place a report of an empirically-based experiment in her portfolio as exemplary work for an undergraduate. It would not be unreasonable for faculty to suggest what classes of products would generate compelling evidence of good or excellent work. Moreover, in order to employ portfolios (or any assessment technique for that matter), faculty need to have established and communicated learning objectives developed at the departmental level.

The second important component of the definition suggests that the selections are made by the student. This means that sometime during their education, students would have to develop criteria and expertise to evaluate their own work. In this light, Stemmer (1993) relates five of the six major premises underlying the use of portfolios to include: (1) Is learner-centered and learner-directed; (2) Is a developmental tool to help the learner set goals and expectations for performance; and (3) Is an instrument that provides a means for the learner to become self-aware and capable of gathering stronger evidence of skills (4); Is a basis for documenting and planning lifelong learning; and (5) constitutes an integration of career planning, counseling, curriculum, instruction and assessment activity.

Finally, the definition of portfolios suggests that selections might come from outside the formal curriculum. For example, a psychology major might list volunteer work from a HeadStart program as part of her portfolio. This work would not only be relevant with regard to the values inculcated by the institution for the purpose of service learning, but the choice itself would be related to the major. Stemmer (1993) reiterates this when he states that the sixth premise of using electronic portfolios is (6) to be inclusive of the entire program.

Shermis & Barrera (2002) document the advantages and disadvantages of portfolios. The major advantage is that portfolios, if well-implemented, prompt students to become self-assessors and generally asks them to articulate why their artifacts are good (or not). Students often "buy in" to portfolios because they can use them for job-seeking, advanced education, and other purposes. On the other hand, portfolios are somewhat labor intensive in both their assemblage and scoring.

One mechanism that might be used to address the labor issue of grading portfolios, especially in electronic form, is automated essay scoring—a relatively recent technological development. It holds promise for establishing national norms against which writing performance might be evaluated, formulating developmental norms that would allow an institution

to track changes in student writing quality over time, and incorporating a mechanism for using formative feedback in literacy (writing) instruction (Shermis & Daniels, 2002).

Automated Essay Scoring: What is it?

Automated essay scoring (AES) engines employ computer technology to evaluate and score written prose. Although most research on this technique has involved the English language, models are being developed concurrently for evaluation of other languages (Shermis & Burstein, 2003). Not all writing genre are included in this definition, and indeed, we suspect that certain ones may never be covered (e.g., poetry). Nonetheless, it is estimated that approximately 90% of required writing in a typical college classroom can be evaluated through AES.

In AES grading, rater behavior is used as the ultimate criterion, though at least one system (Intelligent Essay Assessor– Landauer, Laham & Foltz, 2003) evaluates content on the basis of external material. Bennett and Bejar (1998) in criticizing the over-reliance on human ratings as the sole criterion for evaluating computerized assessment performance, claim that such ratings, typically based on a within domain constructed rubric, may ultimately achieve acceptable reliability, but at the cost of external validity. They suggest that three issues must first be addressed in order to maximize the validity of the rating process: First, there is no theory

per se for what constitutes good writing, so using an evaluation scheme in a vein suggested by Messick (1989) is difficult.

Second, it appears as if "good writing" rules are made to be broken. It is only when the writer violates general rules of grammar and syntax that a consensus can be formulated concluding that the writing is less than satisfactory. In this light, even with substantial training and good evaluation rubrics, high reliability of ratings among humans is hard to achieve. Third, even when good reliability among human raters is obtained, it is sometimes for different reasons. The best conclusion that can be reached is that it is hard to get raters to articulate why an essay is good (or bad), but that they can recognize good writing when they see it (Shermis, Koch, Page, Keith, & Harrington, 2002).

Page and Peterson (1995) discuss the use of *proxes* and *trins* as a way to think about the process of emulating rater behavior. *Trins* represent the characteristic dimension of interest such as fluency or grammar whereas *proxes* (taken from approximations) are the observed variables with which the computer works. These are the variables into which a computer parser might classify text (e.g., part of grammar, word length, word meaning, etc.). In social science research, a similar distinction might be made between the use of latent and observed variables.

In terms of its present development, one might think of AES as representing the juncture between cognitive psychology and artificial intelligence. The AES engines, described in the following section, demonstrate that the correlation of technology with human rater behavior. The AES engines, predict as well or better than scores produced by raters, and yields a high degree of construct validity. Explanations as to why it works well are only beginning to emerge as implicit or tacit "trade secrets", and may not correspond well to past research (Shermis & Burstein, 2003). Accordingly, the technology must be viewed "in the making" akin to where microcomputers were in the early 1980's, impressive for the time being, but having the potential for improvement.

The AES Scoring Engines

The first automated essay scorer to be developed was Project Essay Grade (PEG; Page, 1966). Although initial work on PEG started in the 1960's, some practical problems weren't solved until the microcomputer became popular in the late 1980's. Acting upon the rising interest in the topic of automated essay scoring within the assessment field, ETS conducted a blind test of PEG for scoring 1,314 essays produced by students taking the Praxis test, used in evaluating applicants for teacher certification (Page & Petersen, 1995). The results supported the hypothesis that PEG was more accurate

in predicting human ratings up to and including three human judges (Page & Petersen, 1995). In addition, these findings demonstrated that by using automated essay scoring (AES), essays could be graded more quickly, more cost-effectively, and more descriptively as compared to using human judgments (Shermis & Burstein, 2003). In essence, the automated grading of essays proved to be not only more accurate, but also more rapid and economical.

By the same token, past work on PEG has yielded favorable results when studying the traits within an essay (e.g. its style, content, and creativity). One recommended use of such traits according to Page (2003) would be "to apply them ipsatively, i.e., comparing the traits as measured within the student". This type of evaluation would yield information as to what trait a specific student is especially strong in and which they need to improve; proving to be an invaluable tool for the improvement of writing skills.

Since the early 1990's, PEG technology has been modified in several ways. For example, it has since acquired several parsers and dictionaries and it has incorporated special collections/classification schemes (Page, 2003). Also, Shermis, Mzumara, Olson, & Harrington (2001) reported on PEG's first use of a web-based interface for grading student placement test essays. The design consisted of 1200 essays scored holistically

by four different raters. The results were encouraging; human judges correlated .62 percent of the time, while PEG correlated with the judges at .71. In addition, the grading speed of PEG improved to evaluating approximately three essays per second (Shermis et al., 2002). In sum, PEG has resulted in a very efficient and economical project that has radically improved the functionality of automated essay grading throughout the years.

Intellimetric

IntelliMetric, a second type of automated essay scorer, has also been shown to be highly effective. It was first made available to educational agencies in January of 1998 and was the first essay-scoring tool based on artificial intelligence. It leverages artificial intelligence research in 4 primary areas: 1) Machine Learning 2) Natural Language (NLU) 3) Pattern Matching and 4) Heuristics Integration. In doing so, Intellimetric™ is able to analyze the content and structure of written works and thus provide a unique evaluation for each one.

Intellimetric™ relies on Vantage Learning's CogniSearch™ and Quantum Reasoning™ technologies, the specific characteristics associated with each score point are internalized and then applied to subsequent scoring. Interestingly, the scoring engine may be said to "learn" which characteristics raters tend to value highly and those that the raters associate with poor scores.

IntelliMetric™ technology parallels processes of holistic scoring and human raters: e.g. on the one hand, human scorers trained to be prompt-specific, and, on the other, Intellimetric™ is able to create a solution for each stimulus prompt (Elliott, 2003). It is capable of analyzing English into seventy-seven semantic, syntactic, and discourse level features (Elliott, 2003) in five different categories: focus and unity, development and elaboration, organization and structure, sentence structure, mechanics and conventions. These have been extended to other languages including French, Dutch, Portuguese, and Italian.

IntelliMetric™ is based on the merging of artificial intelligence, natural language processing, and statistical technologies. It has been used to score open-ended, essay-type questions in English, Spanish, Hebrew and Bahasa (Elliott, 2003).

IntelliMetric™ uses a multi-stage procedure to score essay-type responses. In the first step, IntelliMetric™ internalizes the known score points of a set of responses. Subsequently, the model is tested against a smaller set of responses with known scores that aides in validation and generalizability of the model. Once these are confirmed, the model is used to score new responses whose scores are unknown. Responses are targeted if they are evaluated to be atypical with regards to the standards

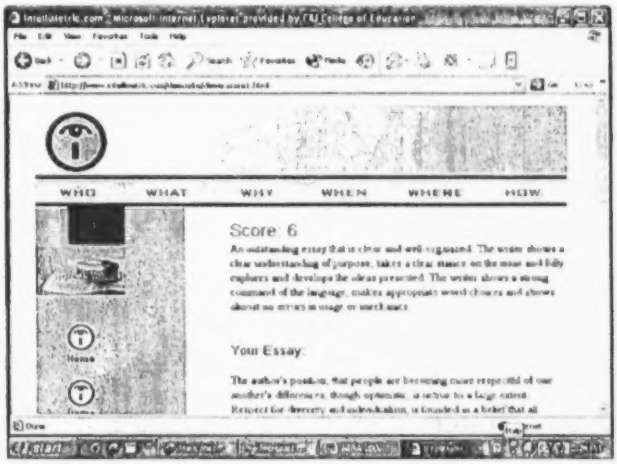
previously set by the essay scoring or by standard American English.

IntelliMetric™ may be applied in either "instructional" or "standardized assessment" modes. As an instructional tool, it provides feedback on a specific student's overall performance. In particular, it provides diagnostic feedback on several dimensions like organization and on analytical dimensions like sentence structure (Elliott, 2003). It permits a student to revise and edit their own essay compositions. The standardized assessment mode is configured to provide for a single student's submission with a holistic score and if need be, feedback on several rhetorical and analytical dimensions (Elliott, 2003).

With regards to the validity of IntelliMetric™, various designs have been employed that fall within three main categories. One is the IntelliMetric™-expert comparison studies, which provides comparisons between IntelliMetric™'s scores and those produced by about two expert raters. The second is the true score studies which uses a large number of expert raters, whose scores are then averaged and used as a proxy for the true score. This true score approximation is then compared to the IntelliMetric™ score and the experts' scores. The third category is that of construct validity studies, in which both the scores produced by IntelliMetric™ and expert raters are compared to other external measures to evaluate whether IntelliMetric™ is

consistent with the expectations for the construct (Elliott, 2003). In sum, IntelliMetric™ has showed greater accuracy in scoring than that of two expert raters (Eliott, 2003). Figure 1 shows a screenshot from the IntelliMetric™ grader.

Figure 1. Screenshot for Intellimetric™ Grader.



Intelligent Essay Assessor

The third essay scoring system in the development of AES is that of the Intelligent Essay Assessor™ (IEA). Based on Latent Semantic Analysis (LSA), it is used for scoring the quality of both conceptual content-based essays and creative narratives. Most importantly, LSA technology provides direct, content-based feedback to instructors or teachers (Landauer, Foltz & Laham, 1998). "LSA provides a representation of an essay's semantic content as a vector (e.g. a set of factor loadings) computed from a set of words contained in the essay. Each vector is compared with another through a cosine, for comparing

BEST COPY AVAILABLE

similarities (Landauer, Laham, & Foltz, 2003). The vector length is defined as the length of each point from the origin.

LSA technology uses three different methods for evaluating both the quality and quantity of knowledge within an essay. They are 1) pre-scored essays of other students; 2) expert model essays and knowledge source materials; 3) internal comparison of an un-scored set of essays (Landauer et al., 2003). These methods provide information regarding the degree to which a specific student's essay has content of the same meaning as that of the comparison texts).

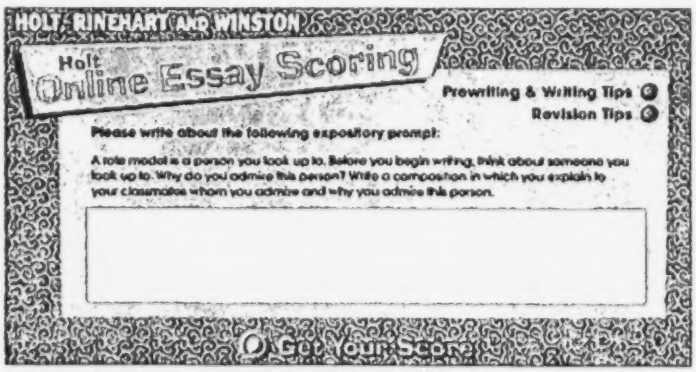
The primary method of evaluation, holistic, compares an essay of unknown quality to a set of pre-scored essays. "In LSA, vectors are used to produce two independent scores, one for the semantic quality of the content, the other for the amount of such content expressed" (Landauer et al., 2003). A quality score is derived by having human raters score a large sample of student essays. Subsequently, each of the human-scored essays is compared with the to-be-scored essays. Then about ten of the pre-scored essays that most resemble the specific target essay are selected. Finally, this target essay is given "the weighted-by-cosine-average human score of those in the similar set" (Landauer et al., 2003).

In particular, the Intelligent Essay Assessor™ has proven to be very useful for not only quick and efficient essay

scoring, but also for detecting plagiarism. Since every essay is compared to every other essay in a given set, if two are found to be similar they are flagged by IEA™ (Landauer et al., 2003). This may prove to be an invaluable tool for educators that do not have the ability, with 150 or more essays to grade, to detect students' plagiarism. Since this form of academic dishonesty is so hard to detect by human scorers, automated essay scoring technology may shed light into a previously illusive concept.

In sum, IEA™'s future consists in expanding beyond the more global assessment of such characteristics like flow and coherence to more specific ones like audience focus and voice (Landauer et al., 2003). Consequently, these improvements may result in the expansion of IEA™ technology for assessment purposes. Figure 2 illustrates a screenshot of IEA™ as implemented in the Holt Online Essay Scoring©.

Figure 2. Screenshot for Holt Online Essay Scoring© (http://www.hrw.com)


HOLT, RINEHART AND WINSTON
Holt
Online Essay Scoring
Prewriting & Writing Tips
Revision Tips
Please write about the following expository prompt:
A role model is a person you look up to. Before you begin writing, think about someone you look up to. Why do you admire this person? Write a composition in which you explain to your classmates whom you admire and why you admire this person.
Get Your Score

BEST COPY AVAILABLE

E-Rater®

The final essay scoring system is e-rater®, developed by the Educational Testing Service (ETS) in 1999 for the operational scoring of the GMAT Analytical Writing Assessment. In use, examinees are assigned an e-rater® score and one human reader score, a process used to score over one-million essays. Studies have shown that e-rater agrees with human reader scores about 97% of the time, thus demonstrating that e-rater technology is a reliable measure of essay scores.

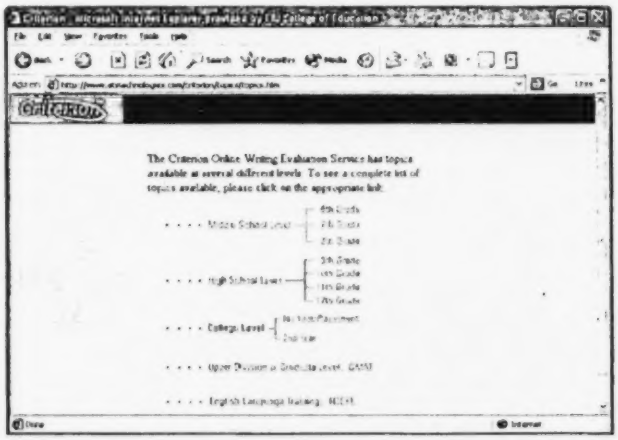
The e-rater® scoring system aims to implement similar features to those used in holistic scoring, yielding a number that represents the essay's quality. Its scoring is based on a six-point scale ranging from "1-deficient" to "6-outstanding". To score on the higher end of the scale, an essay must remain consistent with its topic and have a strong, well-organized argument. In addition, an essay must also consist of a strong syntactic structure and use a diversity of words (Burstein, 2003). E-rater®'s features characterize the essay's syntactic structure, discourse structure, vocabulary usage and lexical complexity. First it builds a statistical model of how these features are related to the scores that human readers have given to a set of training essays and it then uses the model to assign scores to new essays.

Recently, e-rater® has been incorporated with Criterion℠, which is an online, web-based, essay evaluation project of ETS Technologies, a for-profit subsidiary of the Educational Testing Service. Currently, this project is used by institutions for high and low-stakes writing assessments, as well as classroom instruction. Through Criterion, students can write an essay on a number of topics, submit it to e-rater for scoring and view their scores within seconds. In addition, Criterion℠ includes the Critique Writing Analysis Tool which provides students with specific diagnostic feedback concerning the structure and quality of their writing.

In sum, e-rater® scores essays based on a prompt-specific model (Burstein, 2003). Presently, e-rater®, supplemented with the Criterion℠ model, provides diagnostic feedback about grammar, mechanics and style and overall holistic scores. In the near future, supplemental feedback will also include the measures of the quality of the thesis statement and the degree to which the main points of the essay are related to the thesis. Current research in automated essay scoring has indicated that e-rater® performs comparably to human readers at different grade levels (Burstein, 2003). More recent research focuses on the development of more generic, global e-rater® scoring models. Burstein (2003) reported that e-rater® models exist for prompts based on data samples from grades 4 through 12 using national

standards prompts; for undergraduates, using English Proficiency Test (EPT) and PRAXIS prompts; and, for non-native English speakers, using TOEFL prompts. ETS programs, including GMAT, TOEFL, and GRE are currently using e-rater® with Criterion℠ for low-stakes, practice tests (Burstein, 2003). Figure 4 shows a screenshot of a topics list of available Criterion℠ prompts.

Figure 4. Criterion℠ topics list for different grade levels.


CRITERION
The Criterion Online Writing Evaluation Service has topics available at several different levels. To see a complete list of topics available, please click on the appropriate link:

The FIPSE Project

Shermis (2000) has designed a FIPSE-funded project to create national norms for documents found commonly in electronic portfolios. These norms will then be available, for a period of five years, through automated software that could grade the documents via the World Wide Web. Documents to be included in the norming procedure have been drawn from four writing genres: reports of empirical research, technical reports, historical narratives, and works of fiction.

BEST COPY AVAILABLE

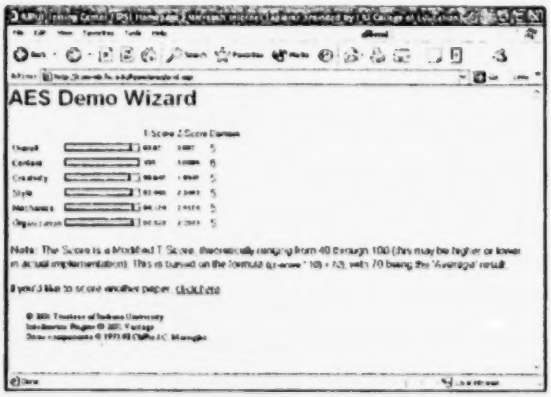
This application is based on previous research with shorter (i.e., less than 500 words) essays in which computers have surpassed both the reliability and validity of human raters. The ultimate criterion in this process are the evaluations of human raters, and the results of regression models of writing based on large numbers of essays and raters. In order to build the statistical models to evaluate the writing, several institutions from across the country, representing a range of Carnegie classifications, have agreed to provide 400-750 documents that are reflective of their current electronic portfolios. Six raters will evaluate each document and provide both holistic and trait ratings.

Vantage Technologies, Inc. has agreed to provide their Intellimetric™ parser for both model building and actual implementation of the project. Post-secondary institutions that are moving towards electronic portfolios could benefit from having access to the comparative information. Moreover, establishing norms would allow an institution to examine writing development of students over time. Finally, the software could be used in a formative manner, allowing students to preview their writing evaluations in order to improve writing or make better document selections.

Because previous work with the Intellimetric™ grading engine placed a heavy emphasis on content, and needed to be

modified to focus on the characteristics of general writing ability, a study was conducted to determine to which it would score as reliably as other engines (Shermis et al., 2002). Moreover, there was a need to test the ability of the Intellimetric™ engine to interact with the project's web-based support mechanisms. The results showed that the modifications to the Intellimetric™ engine resulted in inter-rater agreement coefficients that were as high, and in a few cases, higher than the AES models tested with shorter documents. Moreover, the web-based support mechanisms used for previous work were easily adaptable to the Intellimetric™ engine. So that prospective users might give the software a "tryout", a site has been set up with a demonstration based on a few different models. This web site is located at: http://coeweb.fiu.edu/fipsedemo. Figure 5 shows a screenshot of a writing feedback page from this site.

Figure 5. A screenshot of the FIPSE-sponsored demonstration site.

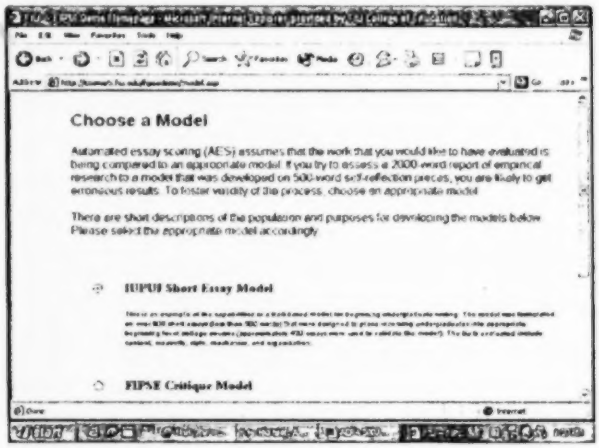


BEST COPY AVAILABLE

Model Building

Vantage Learning has recently completed work on the "critiques" genre and has created an operational model for use. This can be accessed by choosing between either the short essay or critique models as shown in Figure 6.

Figure 6. Choosing the "critique" model at the FISPE AES site.



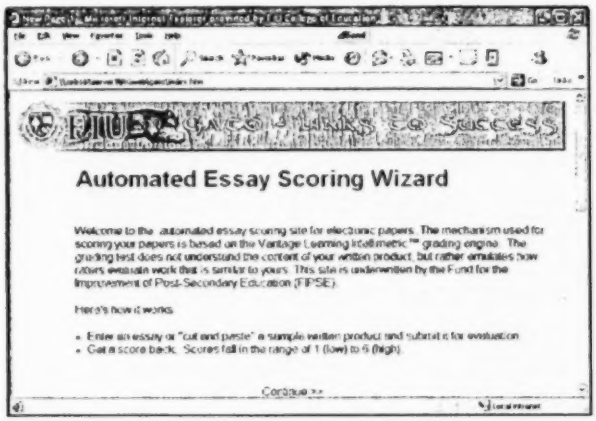
Work on the "self-reflective" writing model is projected to be complete by mid-April.

In early spring, the FIPSE-sponsored automated essay scoring technology was incorporated into "GATO" website at Florida International University as part of a suite of tools used to support undergraduate education efforts. This aspect of the website is designed to help students in writing courses obtain feedback on their drafts prior to submitting it to the instructor. A study is presently being planned to evaluate the

BEST COPY AVAILABLE

application of automated essay scoring in writing performance. Figure 7 illustrates the gateway to this writing resource.

Figure 7. A screenshot of GATO, a set of web site tools for general education (including writing).


Automated Essay Scoring Wizard
Welcome to the automated essay scoring site for electronic papers. The mechanism used for scoring your papers is based on the Vantage Learning IntelliMetric™ grading engine. The grading tool does not understand the content of your written product, but rather emulates how raters evaluate work that is similar to yours. This site is underwritten by the Fund for the Improvement of Post-Secondary Education (FIPSE).
Here's how it works:
• Enter an essay or "cut and paste" a sample written product and submit it for evaluation
• Get a score back. Scores fall in the range of 1 (low) to 6 (high)
Continue >>

A similar study is being contemplated for three of the five campuses at Miami-Dade Community College which is the largest community college in the U.S. MDCC's current plans are to incorporate feedback from the automated essay scoring engine to determine to what degree such feedback improves writing scores. One possible outcome of this association is the ability to study students where English is a second language.

Yet a third study is near completion with the Miami-Dade County Public Schools that has exactly the same objective, but for a slightly different population. In this study, half the students in a large urban high school were taught 10th grade writing with the support of automated essay scoring and half the students did not have access to the technology. Later this year

BEST COPY AVAILABLE

we will be able to evaluate writing performance against the Florida version of their statewide accountability measure, FCAT (Chodorow, 2002).

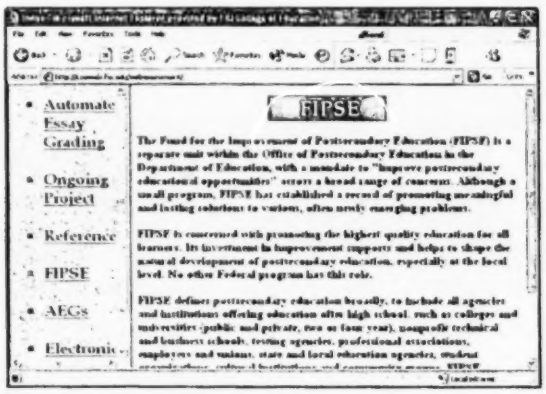
Dissemination Activities

In addition to the demonstration web site, the project has created an informational web site that describes some of the on-going activities associated with automated essay scoring, provides links to all of the major automated essay scorers, and gives references and contacts for those involved in automated essay scoring. This website can be found at:

http://coeweb.fiu.edu/webassessment

A screenshot from the information site is given in Figure 8.

Figure 8. A screenshot of the FIPSE-sponsored informational site.



Conclusions

In this paper, we have provided background information on what automated essay scoring is, a brief review of four popular automated essay scoring engines, and an update to a FIPSE-

BEST COPY AVAILABLE

sponsored project that incorporates automated essay scoring into electronic portfolios, and a hint as to where we see future research in the area.

If this project is successful, then it may simply be a matter of some minor programming to incorporate the AES models described herein as part of a distance learning package (for formative use) or as component of an institutional portfolio that monitors student progress on principles of undergraduate learning (a summative use).

Employing national norms for automated essay grading in this fashion can supplement locally-developed human-administered rubrics that focus on content in the major or indicators for program improvement. AES, as described here, is not meant to preclude assessment by humans, but makes possible a more thorough evaluation of students' written work. This information can be very helpful for improving writing, modifying programs of instruction, or making some global assessment of the state of general education in an institution.

Author Notes

Correspondence concerning this article should be addressed to Mark D. Shermis, Florida International University, ZEB 310 University Park, Miami, FL 33199. Electronic mail may be sent via Internet to MShermis@ FIU.Edu. Research for this project was sponsored by the Fund for the Improvement of Post-Secondary Education (FIPSE Grant # P116B000387A). The opinions expressed in this paper do not necessarily reflect those of FIPSE or the U.S. Department of Education.

References

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Burstein, J. (2003). The E-rater™ Scoring Engine: Automated Essay Scoring With Natural Language Processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach* Mahwah, NJ: Lawrence Erlbaum.

Chodorow, M. (2002, December). E-rater®: An automated system for scoring essays. Paper presented at the annual meeting of the National Institute of the Japanese Language, Toyko, Japan.

Elliot, S. (2003). Intellimetric™: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach*. Mahwah, NJ: Lawrence Erlbaum.

Hatfield, S. (1997, November). *Assessment in the major: Tools and tips for getting started.* Paper presented at the Assessment Conference in Indianapolis, Indianapolis, IN.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998) An introduction to latent semantic analysis. Discourse Processes, 25, 2&3, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: MacMillan.

NWREL. (1999, December). *6+1 Traits™ of Writing Rubic* [web site]. Northwest Educational Research Laboratory. Retrieved, from the World Wide Web: http://www.nwrel.org/eval/pdfs/6plus1traits.pdf

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238-243.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach*. Mahwah, NJ: Lawrence Erlbaum.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(6), 561-566.

Shermis, M. D. (2000). *Automated essay grading for electronic portfolios* (Grant No. P116B000387A). Washington, DC: Fund for the Improvement of Post-Secondary Education.

Shermis, M. D. (in press). Facing off on automated essay scoring. *Assessment Update*.

Shermis, M. D., & Barrera, F. D. (2002). Automated essay scoring for electronic portfolios. *Assessment Update, 14*(4), 1-2.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross disciplinary approach.* Mahwah, NJ: Lawrence Erlbaum.

Shermis, M. D., & Daniels. (2001) Automated essay grading for electronic portfolios. *Assessment Update,* 13 (1), 10.

Shermis, M. D., & Daniels, K. (2002). Web applications in assessment. In T. W. Banta (Ed.). *Building a Scholarship of Assessment.* San Francisco: Jossey-Bass (pps. 148-166).

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62* (1), 5-18.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education, 26*(3), 247-259.

Stemmer, P. (1993, February). *Electronic portfolios: Are going to the very next craze.* Paper presented at the Michigan School Testing Conference, Ann Arbor, MI.

Vantage Learning (2000). *A true score study of Intellimentric™ accuracy for holistic and dimensional scoring of college*

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM035025

I. DOCUMENT IDENTIFICATION:

Title: Assessing Writing Through the Curriculum with Automated Essay Scoring	
Author(s): Mark D. Shermis, Marylou V. Raymat, & Felicia Barrera	
Corporate Source: Paper presented at AERA, Chicago, IL	Publication Date: April 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

X

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Mark D. Shermis	Printed Name/Position/Title: Mark D. Shermis, Assoc. Dean
Organization/Address: College of Education, ECB 310, FIU, Miami, FL 33199	Telephone: 305-348-2092 / FAX: 305-348-2081
	E-Mail Address: mshermis@fiu.edu / Date: 6/4/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)